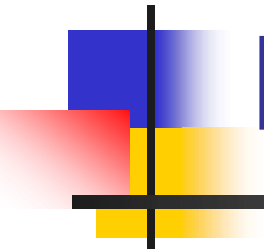
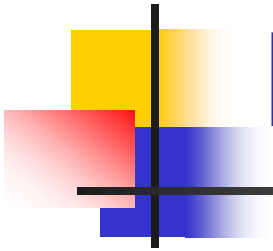


Chapter 1 Introduction to Data Mining

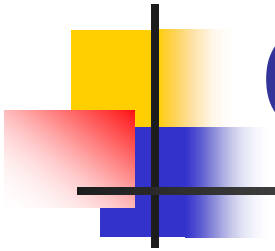
Dr. Bernard Chen

University of Central Arkansas



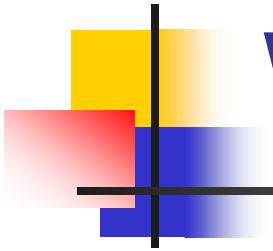
Data Mining Class

- **This class is an introduction** to a young and promising filed called *data mining* and *knowledge discovery from data*



Outline

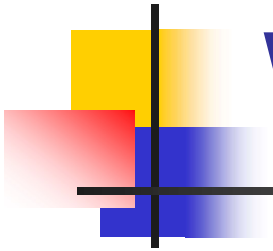
- What Motivated Data Mining?
- So, What Is Data Mining?
- What kind of patterns can we mined?



What Motivated Data Mining?

- Necessity is the mother of invention – Plato

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
  - Major sources of abundant data

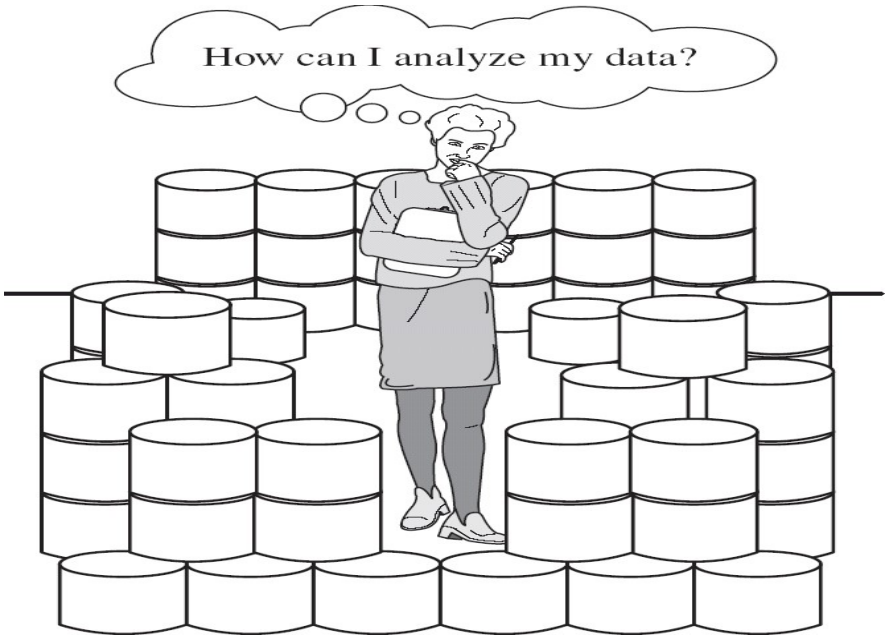


What Motivated Data Mining?

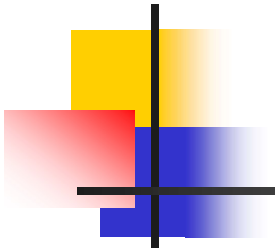
- Data collection and data availability
  - Automated data collection tools, database systems, Web, computerized society
- Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, scientific simulation, …
  - Society and everyone: news, digital cameras, YouTube

What Motivated Data Mining?

- We are drowning in data, but starving for knowledge!

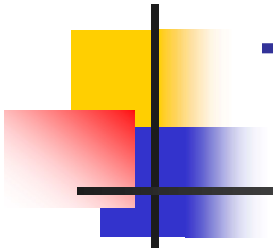


Evolution of Database Technology

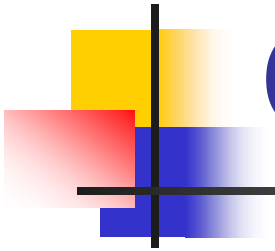


- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

Evolution of Database Technology



- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems
- 2010 to 2020
  - Data Science

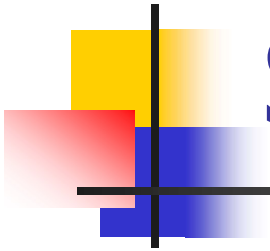


Outline

- What Motivated Data Mining?
- So, What Is Data Mining?
- What kind of patterns can we mined?

So, What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?

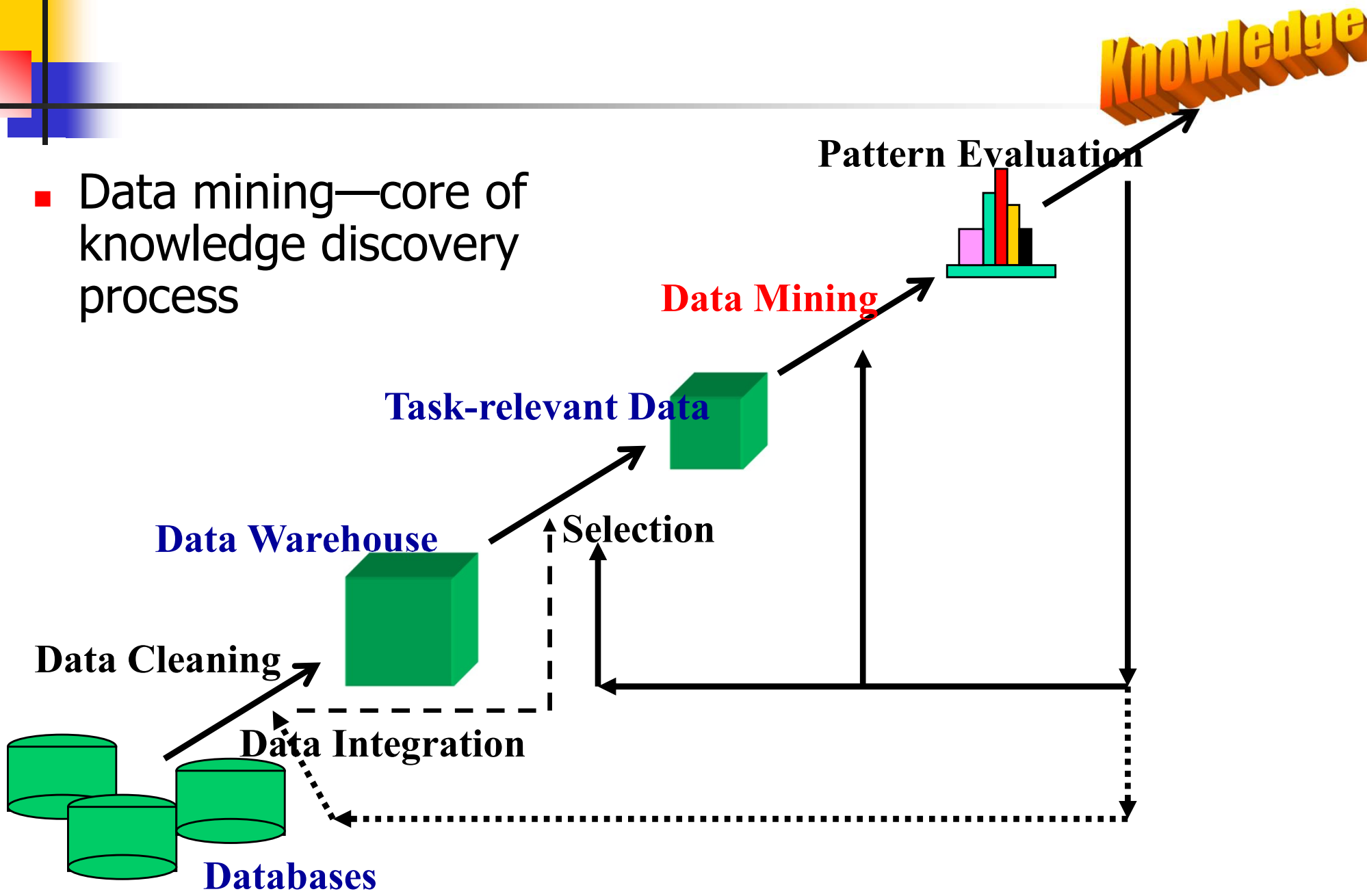


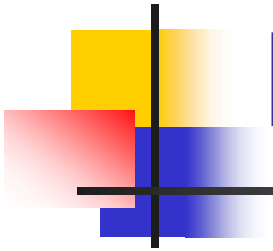
So, What Is Data Mining?

- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Knowledge Discovery (KDD) Process

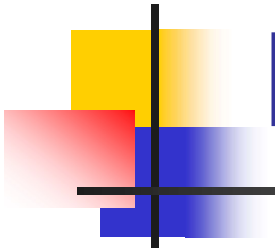
- Data mining—core of knowledge discovery process





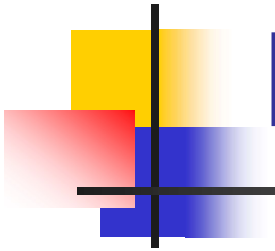
Knowledge Process

1. **Data cleaning** – to remove noise and inconsistent data
2. **Data integration** – to combine multiple source
3. **Data selection** – to retrieve relevant data for analysis
4. **Data transformation** – to transform data into appropriate form for data mining
5. **Data mining**
6. **Evaluation**
7. **Knowledge presentation**



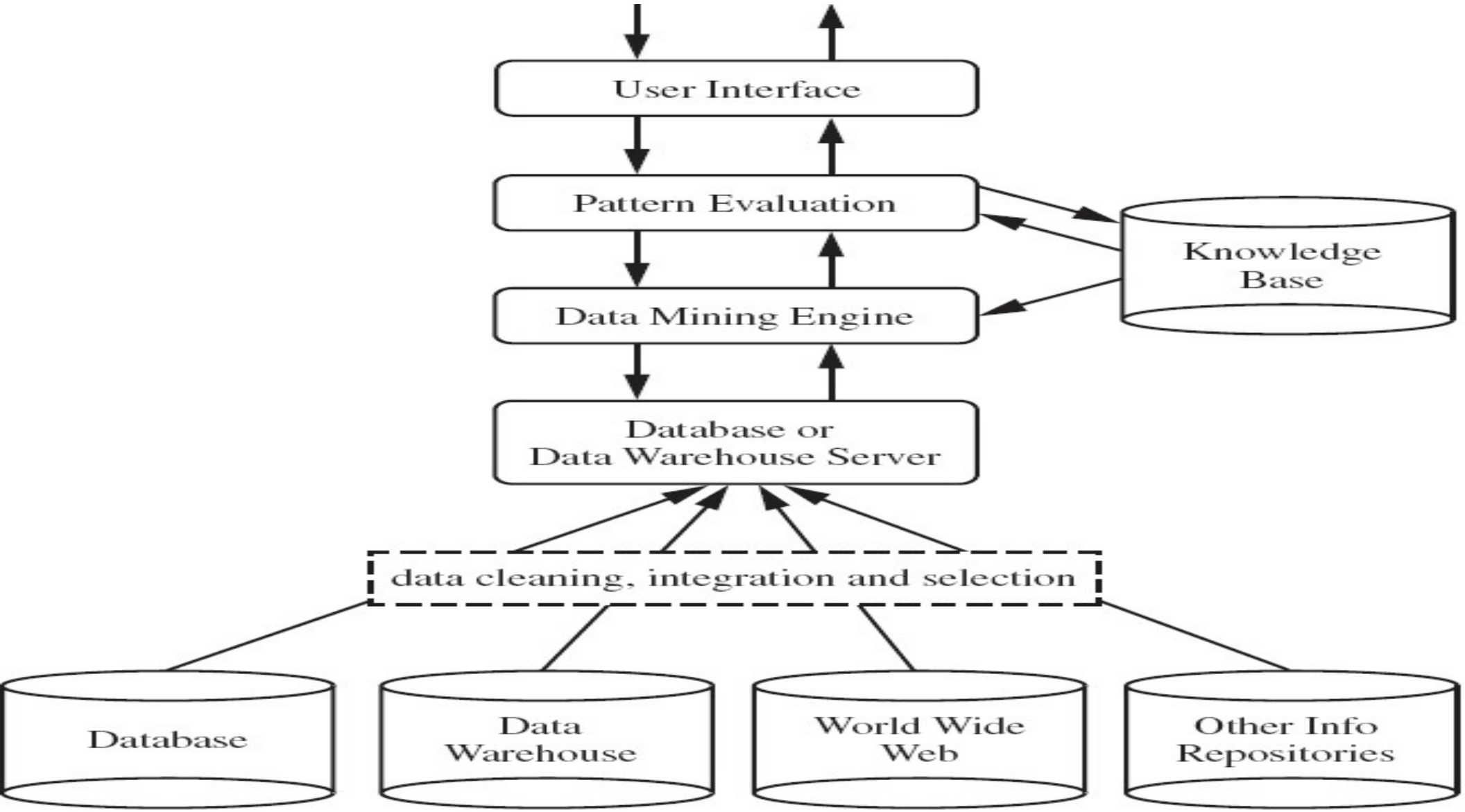
Knowledge Process

- Step 1 to 4 are different forms of data **preprocessing**
- Although data mining is only one step in the entire process, it is an essential one since it uncovers hidden patterns for evaluation

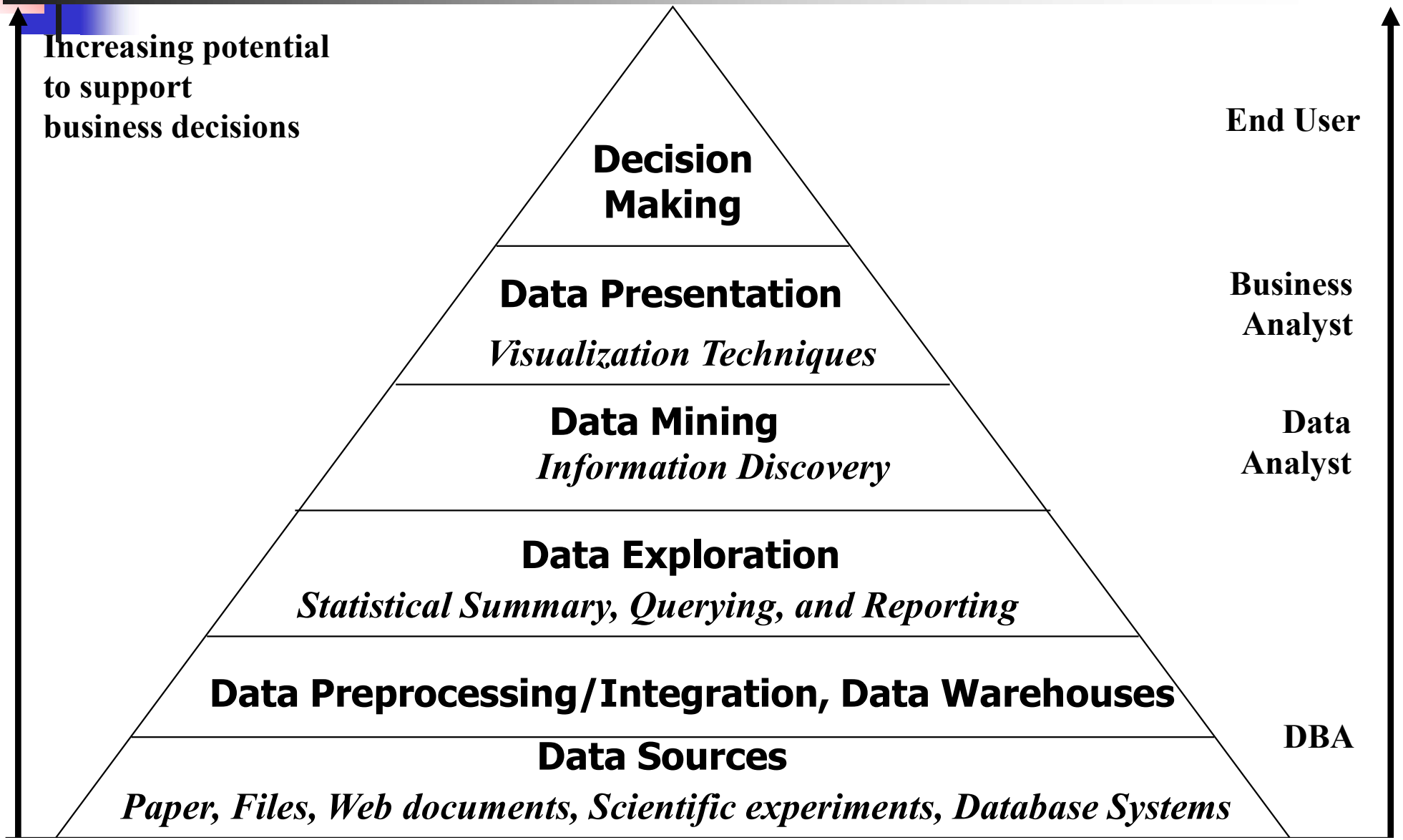


Knowledge Process

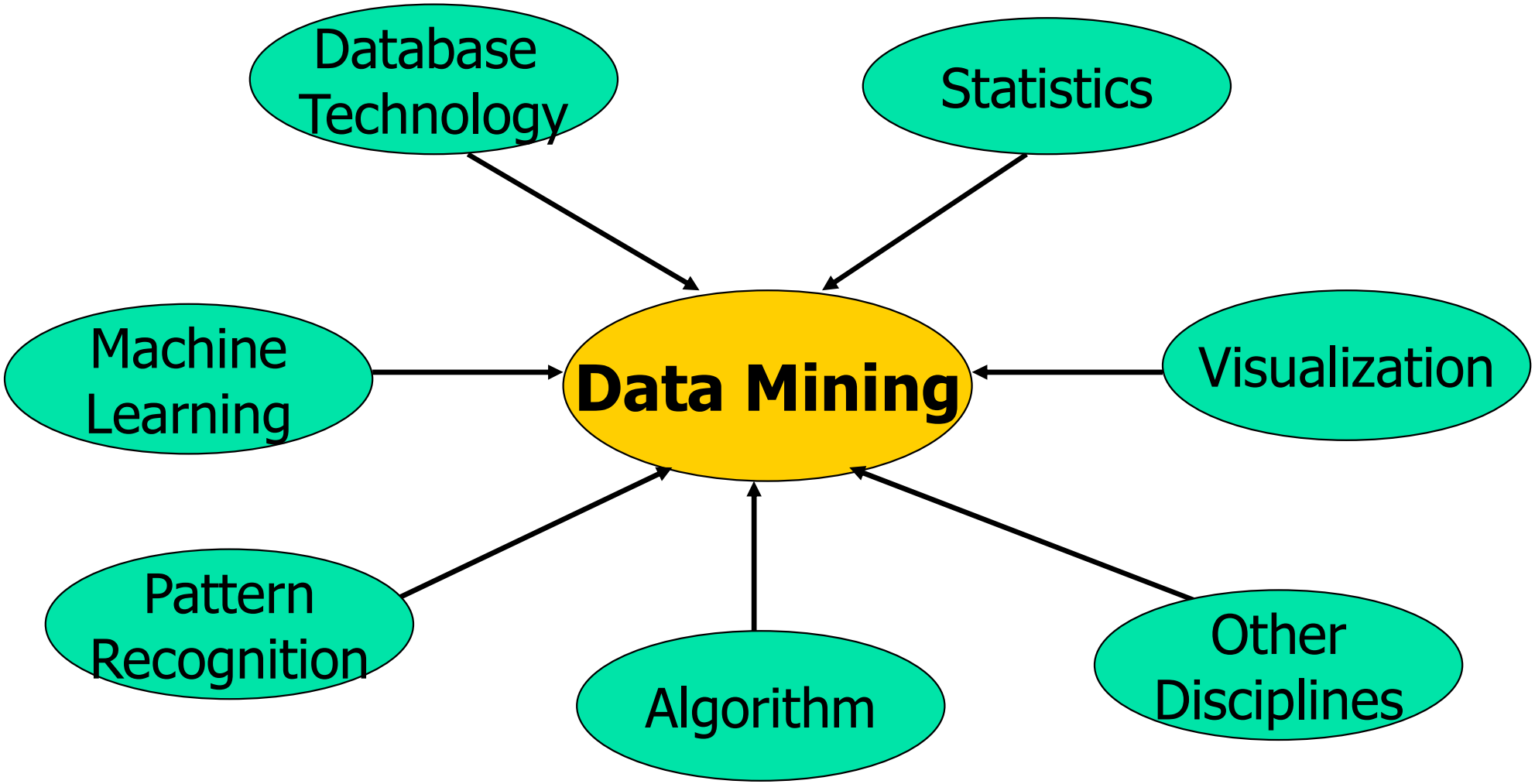
- Based on this view, the architecture of a typical data mining system may have the following major components:
  - Database, data warehouse, world wide web, or other information repository
  - Database or data warehouse server
  - Data mining engine
  - Pattern evaluation model
  - User interface


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

Data Mining and Business Intelligence



Data Mining: Confluence of Multiple Disciplines



Data Mining – on what kind of data?

- Relational Database

Figure 5.6
One possible database state for the COMPANY relational database schema.

EMPLOYEE

Fname	Minit	Lname	Ssn	Bdate	Address	Sex	Salary	Super_ssn	Dno
John	B	Smith	123456789	1965-01-09	731 Fondren, Houston, TX	M	30000	333445555	5
Franklin	T	Wong	333445555	1955-12-08	638 Voss, Houston, TX	M	40000	888665555	5
Alicia	J	Zelaya	999887777	1968-01-19	3321 Castle, Spring, TX	F	25000	987654321	4
Jennifer	S	Wallace	987654321	1941-06-20	291 Berry, Bellaire, TX	F	43000	888665555	4
Ramesh	K	Narayan	666884444	1962-09-15	975 Fire Oak, Humble, TX	M	38000	333445555	5
Joyce	A	English	453453453	1972-07-31	5631 Rice, Houston, TX	F	25000	333445555	5
Ahmad	V	Jabbar	987987987	1969-03-29	980 Dallas, Houston, TX	M	25000	987654321	4
James	E	Borg	888665555	1937-11-10	450 Stone, Houston, TX	M	55000	NULL	1

DEPARTMENT

Dname	Dnumber	Mgr_ssn	Mgr_start_date
Research	5	333445555	1988-05-22
Administration	4	987654321	1995-01-01
Headquarters	1	888665555	1981-06-19

DEPT_LOCATIONS

Dnumber	Dlocation
1	Houston
4	Stafford
5	Bellaire
5	Sugarland
5	Houston

WORKS_ON

Essn	Pno	Hours
123456789	1	32.5
123456789	2	7.5
666884444	3	40.0
453453453	1	20.0
453453453	2	20.0
333445555	2	10.0
333445555	3	10.0
333445555	10	10.0
333445555	20	10.0
999887777	30	30.0
999887777	10	10.0
987987987	10	35.0
987987987	30	5.0
987654321	30	20.0
987654321	20	15.0
888665555	20	NULL

PROJECT

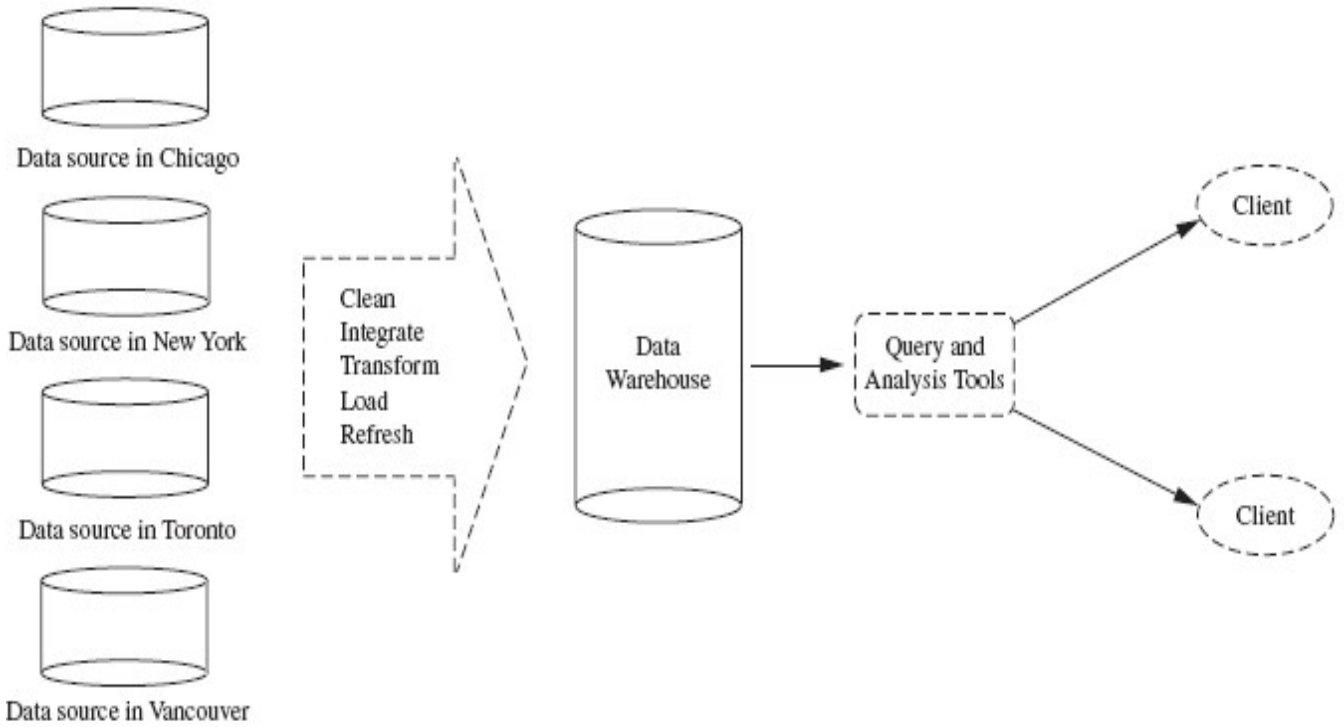
Pname	Pnumber	Plocation	Dnum
ProductX	1	Bellaire	5
ProductY	2	Sugarland	5
ProductZ	3	Houston	5
Computerization	10	Stafford	4
Reorganization	20	Houston	1
Newbenefits	30	Stafford	4

DEPENDENT

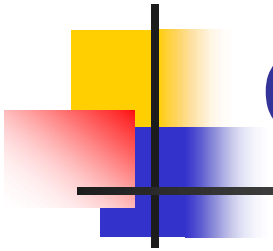
Essn	Dependent_name	Sex	Bdate	Relationship
333445555	Alice	F	1986-04-05	Daughter
333445555	Theodore	M	1983-10-25	Son
333445555	Joy	F	1958-05-03	Spouse
987654321	Abner	M	1942-02-28	Spouse
123456789	Michael	M	1988-01-04	Son
123456789	Alice	F	1988-12-30	Daughter
123456789	Elizabeth	F	1967-05-05	Spouse

Data Mining – on what kind of data?

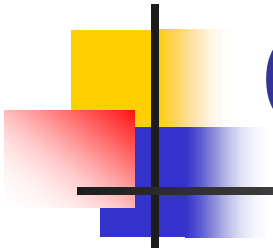
- Data Warehouses



Data Mining – on what kind of data?



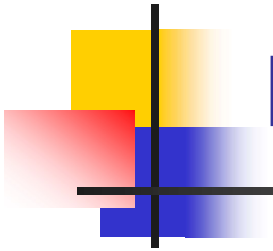
- Transactional Databases
- Advanced data and information systems
  - Object-oriented database
  - Temporal DB, Sequence DB and Time serious DB
  - Spatial DB
  - Text DB and Multimedia DB
  - … and WWW



Outline

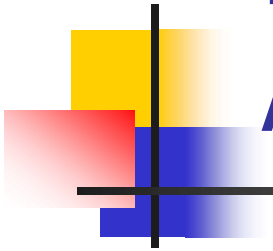
- What Motivated Data Mining?
- So, What Is Data Mining?
- What kind of patterns can we mined?

What kind of patterns can we mined?



- In general, data mining tasks can be classified into two categories: **descriptive** and **predictive**
  - Descriptive mining tasks characterize the general properties of the data in database
  - Predictive mining tasks performs inference on the current data in order to make predictions

Mining frequent patterns, Associations, and Correlations (Ch4)

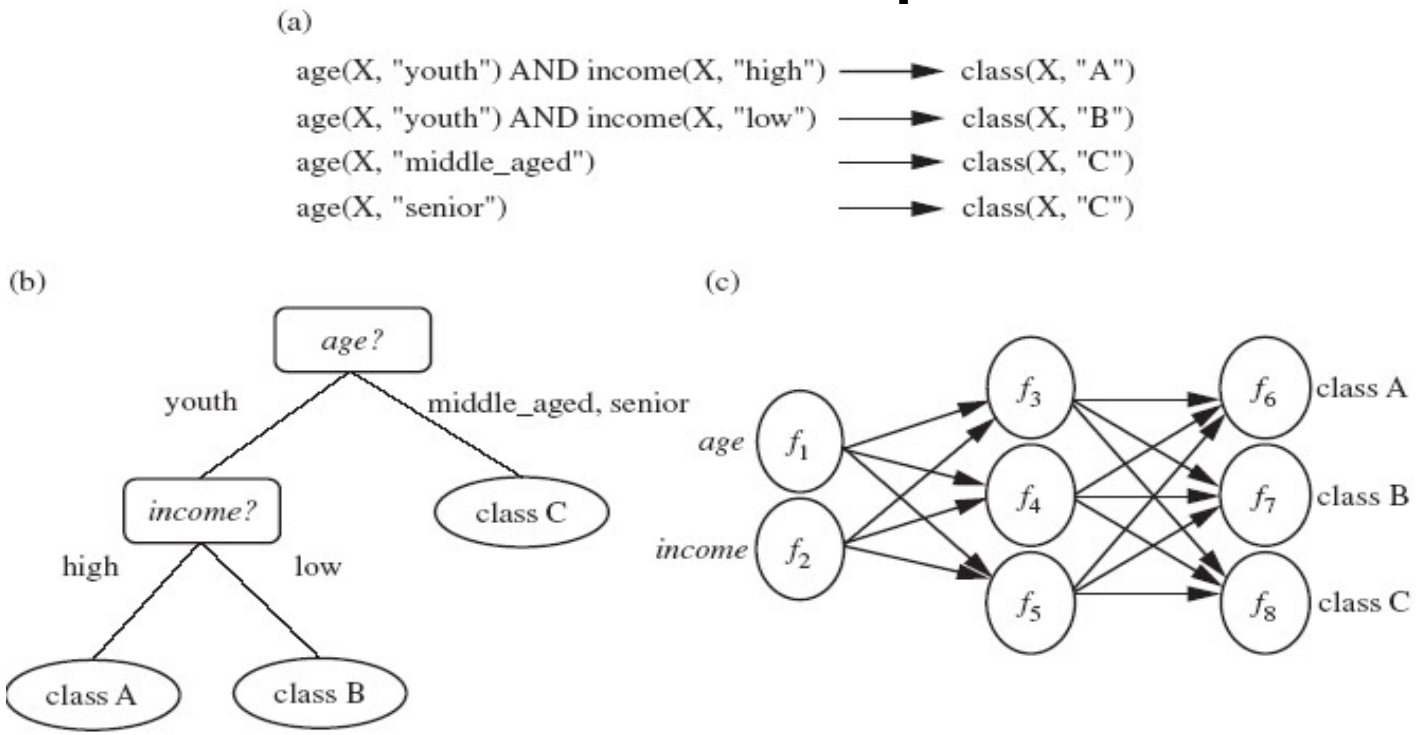


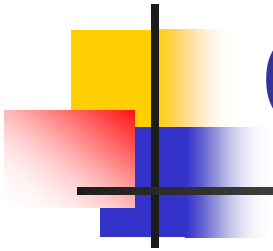
- **Frequent patterns** are patterns that occur frequently in data

- Association analysis:
  - Example: buys(X,”computer”) => buys(X,”software”) [support = 1%, confidence = 50%]

Classification and Prediction (Ch 5)

- **Classification** is the process of finding a MODEL that describes and distinguish data classes or concepts



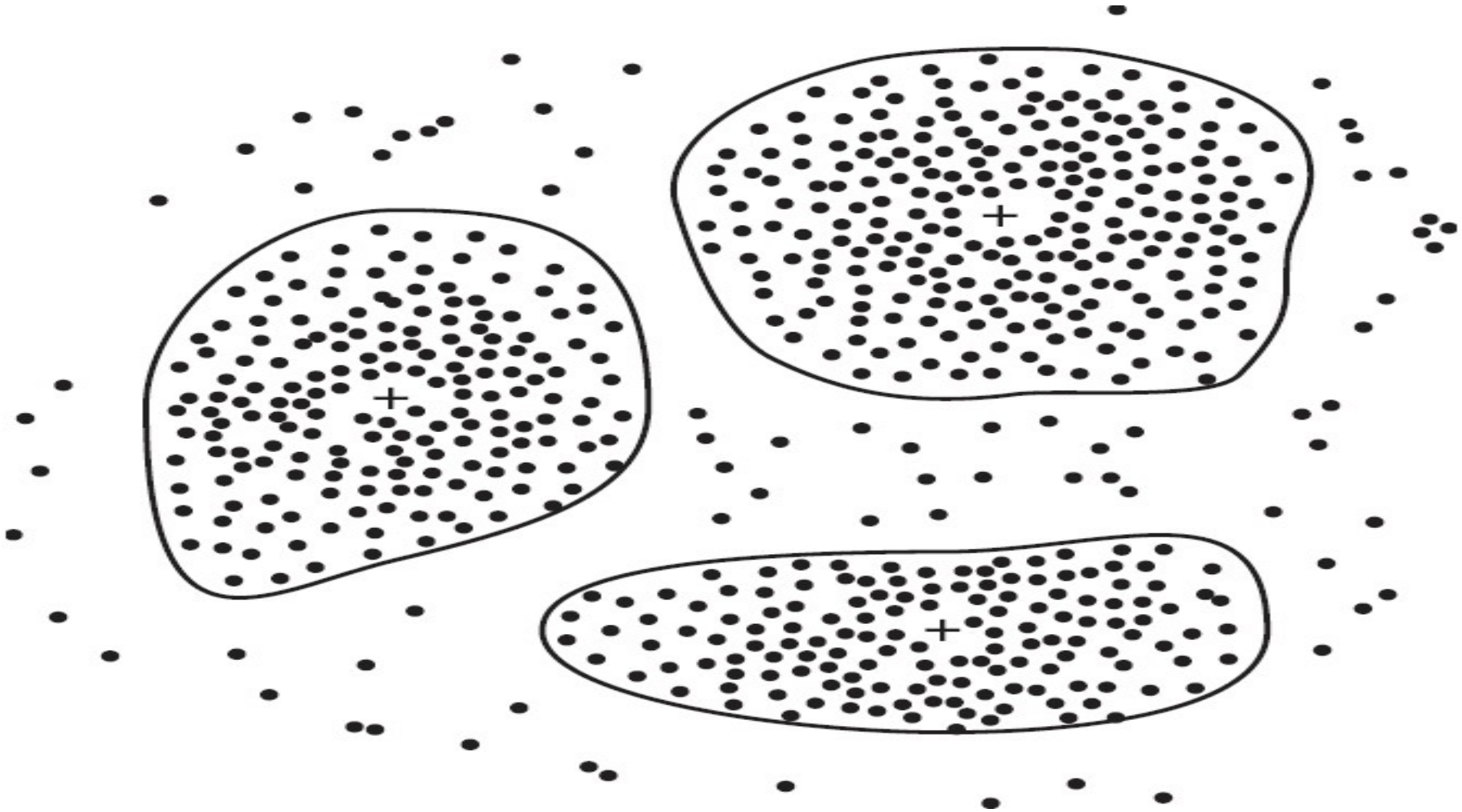


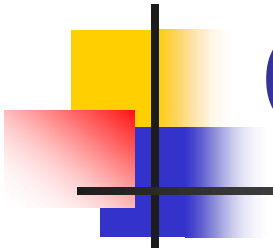
Cluster analysis (Ch 6)

- In general, the class label are not present in the training data simply they are not known to begin with

- The objects are clustered or grouped based on the principle of *maximizing the intra-cluster similarity* and *minimizing the inter-cluster similarity*

Cluster analysis





Outlier Analysis (Ch 7)

- Most data mining methods discard outliers as noise or exceptions.
- However, in some application such as fraud detection, the rare event can be more interesting than regularly occurring ones